Simulating the Lateral Reader for News Trustworthiness Reports with an Iterative Multi-Agent RAG System

Dake Zhang
University of Waterloo
Waterloo, Ontario, Canada

Mark D. Smucker
University of Waterloo
Waterloo, Ontario, Canada

Abstract

Readers of online news often lack the time and domain expertise required to verify unfamiliar claims and sources. Professional fact-checkers address this gap through *lateral reading*, an iterative workflow of asking investigative questions, searching for external evidence, and synthesizing findings with attribution. We present an iterative multi-agent Retrieval-Augmented Generation (RAG) system that operationalizes this workflow for the TREC 2025 DRAGUN Track. Given a news article, specialized agents (1) generate investigative queries, (2) retrieve and filter evidence from the MS MARCO V2.1 Segmented Corpus using a three-stage retriever (BM25+RM3, cross-encoder reranking, and LLM-based selection), and (3) apply an information-sufficiency evaluator that decides whether additional searching is required before writing. The final report generator produces a 250-word trustworthiness report grounded in retrieved segments, guided by automatically generated critical investigative questions. On the official DRAGUN rubric-based evaluation with 30 news articles, our system using `GPT-4.1` ranked first on report generation quality, achieving the highest mean supportive score (0.230) with low contradiction (0.013).

CCS Concepts

• **Information systems** → **Retrieval tasks and goals**; **Web searching and information discovery**; • **Computing methodologies** → **Natural language generation**; *Multi-agent systems.*

Keywords

Text REtrieval Conference; News Trustworthiness; Multi-Agent Systems; Retrieval-Augmented Generation

ACM Reference Format:

Dake Zhang and Mark D. Smucker. 2026. Simulating the Lateral Reader for News Trustworthiness Reports with an Iterative Multi-Agent RAG System. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3805712.3809973

1 Introduction

Online news is one of the primary channels through which the public receives information about science, health, politics, and other topics of interest. Its rapid proliferation has enabled low-quality and deceptive content to propagate at scale, often faster than truthful content and to a broader audience [18]. Misinformation can erode trust, increase polarization, and threaten societal stability [13, 17], motivating research on tools that help people evaluate the trustworthiness of what they read [2, 3, 7, 10].

For online news readers, the bottleneck is not access to information but the ability to make informed judgments, whether due to insufficient domain knowledge [15], limited cognitive resources for conducting research [11], or exposure to intentionally misleading content [1]. Professional fact-checkers overcome these barriers through *lateral reading*: rather than evaluating an article by staying on the page (*vertical reading*), they leave it after a quick scan to check who published it, verify claims against primary sources, and compare how other credible outlets cover the story [20]. The comparative study by Wineburg and McGrew [20] shows that this strategy yields more accurate judgments and exposes systematic failure modes of page-bound evaluation. Automating this proven effective strategy could benefit everyday readers who lack the time, domain expertise, or investigative skills to perform such in-depth trustworthiness assessments on their own.

Prior work on automated fact-checking, however, has largely decomposed the problem into extracting claims, gathering evidence, and predicting veracity [5], with benchmarks such as FEVER [16] and LIAR [19] driving progress on entailment-style verification of isolated claims. This formulation does not fully capture news trustworthiness [12]: an article may contain accurate facts yet still mislead by leaving out key details, lacking context, or relying on dubious sources. Readers need not only answers to specific factual questions but also guidance on what questions to ask (i.e., what aspects of an article need scrutiny), and trustworthiness assessments grounded in verifiable evidence that they can examine.

The building blocks for such an automated system are now feasible. Retrieval-Augmented Generation (RAG) [8] can reduce Large Language Model (LLM) hallucination by grounding generation in retrieved evidence [6], while multi-agent frameworks such as AutoGen [21] enable LLMs' iterative planning, execution, and reflection [4]. ReadProbe [22] was one of the early systems to combine retrieval-augmented LLMs with lateral reading, using LLMs for question (query) generation and answer synthesis over Bing search results. However, it used a single-pass RAG pipeline and was presented as a demonstration without evaluation.

We present an iterative multi-agent RAG pipeline that addresses both limitations. Simulating the lateral reader, specialized agents generate investigative queries, retrieve and filter evidence through a three-stage process (BM25+RM3, cross-encoder reranking, and LLM-based selection), and an information-sufficiency evaluator decides whether to loop back for additional evidence or proceed to writing. This iterative design allows the system to adaptively

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809973

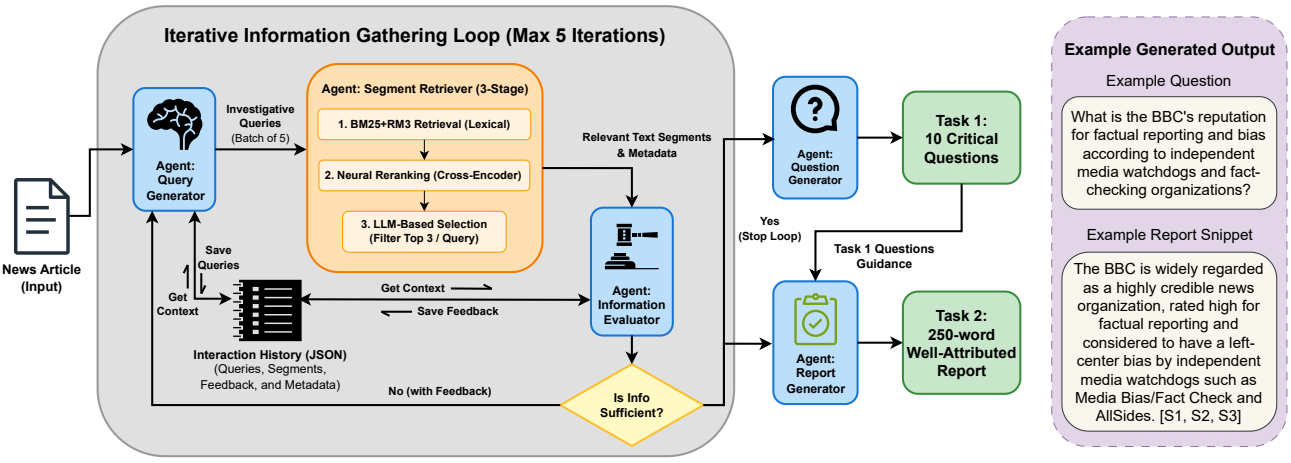

Figure 1: Architecture of our iterative multi-agent RAG system for news trustworthiness assessment. Agents cycle through information gathering and evaluation until deciding that sufficient information has been collected for the final tasks.

fill gaps identified in earlier rounds of investigation, rather than committing to a single pass.

Our system was evaluated on the TREC 2025 DRAGUN (**D**etection, **R**etrieval, and **A**ugmented **G**eneration for **U**nderstanding **N**ews) Track [24], which, building on the TREC 2024 Lateral Reading Track [23], provides the first shared benchmark for this problem: systems need to produce a 250-word report that covers what a reader should know to better assess the trustworthiness of a given news article, grounded in the MS MARCO V2.1 Segmented Corpus[1]. On the official rubric-based human evaluation with 30 news articles, our system using `GPT-4.1` achieved the highest mean supportive score (0.230) among all 28 runs from 8 teams while maintaining low contradiction (0.013).

We release the full implementation and intermediate artifacts to support future system development and analysis: https://github.com/trec-dragun/2025-starter-kit.

2 TREC 2025 DRAGUN Track

The TREC 2025 DRAGUN Track [24] provides a shared benchmark for systems that assist readers in assessing news trustworthiness through lateral-reading-style investigation. The track selected 30 news articles from the MS MARCO V2.1 Document Corpus, covering controversial or context-demanding topics published by 28 media sources with diverse political perspectives.

The track defines two complementary tasks. Task 1 (Question Generation) requires producing a ranked list of ten investigative questions that a careful reader should consider when assessing the article's trustworthiness. Task 2 (Report Generation), the main task, requires producing a well-attributed report of up to 250 words that provides essential background and context for evaluating the article. The report must be grounded in the MS MARCO V2.1 Segmented Corpus, which contains approximately 114 million segments derived from 11 million web documents; each segment is a sliding window of 10 sentences with a stride of 5. Due to space constraints, this paper focuses on Task 2; our system also produces the Task 1 questions and uses them to guide report composition (Section 3).

For evaluation, TREC assessors independently investigated each article's trustworthiness via open-web research (e.g., source credibility, corroboration of key claims, and missing context) and constructed per-article rubrics. Each rubric consists of critical questions $q \in Q$, each paired with one or more expected short answers A_q and an importance weight $w_q \in \{4, 2, 1\}$ (*have to know* = 4, *good to know* = 2, *nice to know* = 1). Assessors treated each rubric answer as a checklist item and labeled the report's relationship to each answer $a \in A_q$ as *supports*, *partial*, *contradicts*, or *none*. Define $s(\cdot) = 1$ for *supports*, 0.5 for *partial*, else 0. The official *supportive* score is an importance-weighted average:

$$\text{Supportive} = \frac{1}{\sum_{q \in Q} w_q} \sum_{q \in Q} w_q \cdot \frac{1}{|A_q|} \sum_{a \in A_q} s(L_{q,a})$$

The *contradictory* score replaces $s(\cdot)$ with $c(\cdot) = 1$ for *contradicts*, else 0. Higher supportive and lower contradictory scores indicate better alignment with assessor findings.

3 System Implementation

Our system is designed as a multi-agent iterative pipeline that gathers information from the specified web corpus and decides when enough evidence has been collected to produce the final outputs. The system consists of several components (agents), each guided by a specialized prompt. Throughout the process, the system maintains a JSON-formatted interaction history of the news article, queries, retrieved text segments, and self-reflection feedback, which each component can consult for context. Figure 1 illustrates the overall architecture of the pipeline.

3.1 Query Generator

As the first agent in the pipeline, given an input news article, the query generator produces a set of search queries that a fact-checker might pose. In the initial iteration, the system prompt instructs the LLM to produce five queries, each accompanied by a rationale,

[1] https://trec-rag.github.io/annoucements/2025-rag25-corpus/#-ms-marco-v21-segmented-corpus

targeting different facets of the article's trustworthiness, e.g., investigating the source's reputation, verifying key claims, tracing quoted facts, checking alternative viewpoints, etc. If the information retrieved later is deemed insufficient by the evaluator (Section 3.3), the query generator is invoked again in subsequent iterations to generate additional queries (again in batches of five). In those later rounds, the agent's input is augmented with context about the previous queries, retrieved segments, and feedback on what information is still lacking. This way, the new queries are informed by what has already been found or missed. This module aims for broad coverage of investigative angles, starting from general and then filling gaps based on the feedback loop from the evaluator.

3.2 Segment Retriever

The segment retriever acts as a search engine to fetch relevant text segments for each query from the query generator. It implements a three-stage retrieval process to progressively narrow down results and extract the most useful segments from the MS MARCO V2.1 Segmented Corpus.

3.2.1 Stage 1: BM25+RM3 Retrieval. For a given query, we first perform lexical retrieval using Pyserini's implementation of BM25 with RM3 pseudo-relevance feedback [9]. We configure BM25 with parameters $k_1 = 0.9$ and $b = 0.4$, and apply RM3 with 10 feedback terms, 10 feedback documents, and an original query weight of 0.5. This efficient lexical stage reduces the search space to a list of 1,000 candidate segments from the index before neural reranking, avoiding the computational overhead of embedding the full 114-million-segment corpus.

3.2.2 Stage 2: Neural Reranking. Next, the candidates are reranked using a cross-encoder model from the Sentence-Transformers library [14]. Specifically, we use `ms-marco-MiniLM-L6-v2`[2], a transformer model fine-tuned on the MS MARCO passage ranking task. For each of the top 1,000 segments, the model examines the query and the segment (title + content) together and produces a relevance score. We then sort the candidates by this neural score and retain the top 20 segments for the final stage. With a richer semantic understanding, this reranking step prioritizes segments that are contextually relevant, even if the wording does not exactly match the query.

3.2.3 Stage 3: LLM-Based Selection. In the final retrieval stage, we take the top 20 segments from Stage 2 and present them (along with the original news article and the query) to an LLM agent with a prompt asking it to select the three most relevant segments for answering the query, ideally from diverse sources. If fewer than three segments are deemed relevant, it selects only those that meet the bar. This LLM-driven filtering helps the pipeline focus on a small set of high-quality pieces of evidence for each query, limiting the inclusion of less relevant material that could dilute the context.

3.3 Information Evaluator

The information evaluator is the agent responsible for deciding whether the system has gathered enough evidence to proceed to final output generation. After each batch of queries and their retrieved segments, the information evaluator LLM examines all information collected so far, including the original news article, the list of queries issued, and the content of the top retrieved segments for those queries. It is instructed to be skeptical and to assess the completeness of the investigation with scrutiny. Essentially, it asks: *Do we now have sufficient information to judge the article's trustworthiness, or should the system dig deeper?*

If the evaluator determines that important questions remain unanswered or key aspects have not been covered, it will output a negative verdict (that information is not yet sufficient) along with a detailed feedback message explaining what is missing or what lines of inquiry should be pursued next. In response, the pipeline loops back to the query generator, providing it with the feedback and the context of past queries with retrieved segments so it can formulate new queries targeting those gaps.

On the other hand, if the evaluator is satisfied that the collected evidence is adequate and no major blind spots remain, it will issue a positive signal indicating that the system can stop searching. This decision triggers the end of the iterative loop, which is capped at five iterations in our implementation for budget control. Thus, through this component, the system achieves an adaptive, iterative search process, i.e., continuing to retrieve information until a stopping criterion is met.

3.4 Question Generator

Once the information-gathering loop terminates, i.e., the information evaluator deems the collected evidence sufficient, the pipeline moves to generating critical investigative questions that will guide report composition. We prompt the LLM with the compiled knowledge base, including the original news article text and all the relevant segments collected during the retrieval iterations, and ask it to formulate the ten most important questions a reader should consider about the article's trustworthiness. The questions are intended to cover the central issues revealed by the evidence, such as the source's credibility, verification of specific claims, presence of bias or counterpoints, etc., essentially summarizing "*What should a skeptical reader ask?*" These questions then serve as a structured outline for the report generator (Section 3.5), guiding the final report to address the most critical trustworthiness dimensions identified during the investigation. The generated questions also constitute our submission for DRAGUN Task 1 (Question Generation).

3.5 Report Generator

The final component is the report generator, which uses a prompt to guide the LLM to act as a professional fact-checker who composes a well-attributed report about the article's trustworthiness. The input to this module includes the news article and the full set of evidence segments retrieved. In addition, we incorporate the generated questions (Section 3.4) as guidance. The prompt encourages the LLM to address those important questions within the report. The report generator's output is a concise narrative that provides relevant background and context to help a reader evaluate the article, within the 250-word limit from the DRAGUN track guidelines. Additionally, the prompt instructs the model to include attribution for each claim by citing the IDs of the supporting segments.

[2]https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

Table 1: Average supportive and contradictory scores across 30 topics for report generation runs, ranked by the primary metric: supportive score. Due to space constraints, we present only the top-ranked runs and the lowest-scoring run.

Run ID	Supportive	Contradictory
our-system-gpt-4.1	0.230	0.013
team-1-run-1	0.218	0.005
team-1-run-2	0.212	0.014
team-1-run-3	0.201	0.014
team-2-run-1	0.173	0.009
team-2-run-2	0.170	0.011
team-2-run-3	0.166	0.025
team-1-run-4	0.158	0.013
team-2-run-4	0.158	0.013
team-2-run-5	0.157	0.018
our-system-gpt-oss-120b	0.150	0.018
team-1-run-5	0.139	0.007
team-3-run-1	0.132	0.008
...	...	...
team-4-run-1	0.005	0.002

4 Results and Discussion

We ran the system with two LLMs during the track participation period (summer 2025): the closed-source `GPT-4.1`[3] (accessed through Azure OpenAI Service; averaged $0.50 and 2 minutes per article) and the open-weight model `gpt-oss-120b`[4] (served locally via vLLM[5]; averaged 3 minutes per article on a single NVIDIA H100 GPU from a university-managed cloud server). Neither model was fine-tuned; all components used the models through API calls with task-specific prompts only. In the DRAGUN track overview paper [24], our `GPT-4.1` and `gpt-oss-120b` runs were submitted under the names `dragun-organizers-starter-kit-task-2` and `organizer-gpt-oss-t2`, respectively. For clarity, we rename them here as `our-system-gpt-4.1` and `our-system-gpt-oss-120b`, and abbreviate other run names in Table 1 due to space constraints. We refer the reader to the overview paper for the full ranking of submitted runs and descriptions of other participants' approaches. We report and discuss the official evaluation results below.

Table 1 shows that `our-system-gpt-4.1` achieves the highest mean supportive score (0.230) while maintaining a low contradictory score (0.013), ranking first among all submitted systems (28 runs from 8 teams) on the primary metric: supportive score. Although the supportive-score differences between the top four runs are modest, our `GPT-4.1`-based system is significantly better than the fifth run through a paired two-tailed Student's t-test ($p < 0.05$). Meanwhile, the gap between our `GPT-4.1`-based system and our open-model variant is also significant (0.230 vs. 0.150 supportive, $p < 0.05$), indicating that stronger model capability is important for finding and synthesizing evidence into a well-attributed report.

Figure 2 shows substantial variance across topics: on the easy topics, `our-system-gpt-4.1` achieves supportive scores around 0.5, while on the hardest ones, neither system surpasses 0.05. But

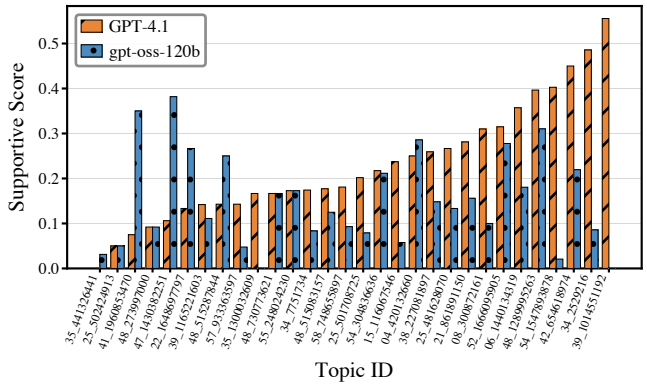

Figure 2: Per-topic supportive scores of our two runs. Topics are sorted by the `GPT-4.1` supportive score (the left bar for each topic) in ascending order. Topic IDs' common prefix "msmarco_v2.1_doc_" is truncated for brevity.

`GPT-4.1` consistently outperforms or matches `gpt-oss-120b` on the majority of topics. Given the high topic variance, we counted per-topic wins, where a system "wins" a topic if it achieves the highest supportive score (with ties counted as shared wins). The `our-system-gpt-4.1` run leads on 12 out of the 30 articles, followed by the second-best run `team-1-run-1` on 8 articles, and the third-best `team-1-run-2` on 3 articles.

An analysis of the information-sufficiency evaluator's behavior in our `GPT-4.1` run reveals a bimodal pattern: for 18 of 30 articles (60%), the evaluator determined that a single retrieval round provided sufficient evidence, while for 9 articles (30%), it exhausted all five permitted rounds. Only 3 articles required an intermediate number of rounds. This suggests that the evaluator makes a largely binary judgment (either the initial evidence is adequate or the topic demands extensive investigation) and that the iterative mechanism is most valuable for the harder subset of articles where straightforward retrieval falls short.

In manual inspection, we observed two issues. First, evidence freshness can cause misalignment: assessor rubrics were written using live web evidence in 2025, while the MS MARCO V2.1 Segmented Corpus predates 2022, which can yield contradictions on time-sensitive facts (e.g., ownership changes or evolving reputations). Future evaluations might benefit from considering temporal alignment between the grounding corpus and rubric construction. Second, our report generator sometimes uses the limited word budget to explicitly state that information was not found for certain investigative questions. In the current evaluation setting, such disclaimers rarely increase supportive score unless the absence itself is diagnostic (e.g., lack of independent corroboration or missing disclosures). A more effective strategy is to instruct the report generator to spend words on findings that can be directly supported by retrieved segments.

Looking ahead, we see three directions for improvement. First, reinforcement learning from assessor rubrics could directly optimize LLM behavior toward the lateral reading standard. Second, the general-purpose LLM powering all agents could be replaced with smaller, task-specialized models (e.g., a dedicated relevance

[3]https://openai.com/index/gpt-4-1/
[4]https://openai.com/index/introducing-gpt-oss/
[5]https://docs.vllm.ai/en/latest/

judge for segment selection or an information-sufficiency evaluator calibrated to fact-checker standards), which could improve both efficiency and component-level accuracy. Third, systematic error analysis comparing LLM-generated reports against human rubrics could pinpoint recurring failure modes (e.g., missing investigation angles) and guide future alignment efforts.

5 Conclusion

We presented an iterative, multi-agent RAG pipeline that operationalizes *lateral reading* for news trustworthiness assessment by looping between investigative query generation, evidence retrieval and filtering, and an explicit information-sufficiency decision before writing. The system produces a concise report grounded in retrieved segments and leverages generated critical questions as guidance for report composition. On the TREC 2025 DRAGUN Track's report generation task, our `GPT-4.1`-based run achieved the highest mean supportive score (0.230) across 30 news articles while maintaining low contradiction (0.013), demonstrating that evaluator-guided iterative retrieval can produce top-performing grounded trustworthiness reports under the 250-word budget. Because the DRAGUN evaluation relies on human assessors applying rubrics to system reports, post-hoc ablation with official scores is not feasible. Component-level analysis with automatic proxy evaluation remains an important direction for future work.

Acknowledgments

This research was supported by Microsoft and by the Natural Sciences and Engineering Research Council of Canada (NSERC) under grant ALLRP/597573-24.

References

[1] Carlos Carrasco-Farré. 2022. The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications* 9, 162 (2022), 1–18. doi:10.1057/s41599-022-01174-9

[2] Michael Chan, Jingjing Yi, Cristian Vaccari, and Masahiro Yamamoto. 2025. A cross-national examination of the effects of accuracy nudges and content veracity labels on belief in and sharing of misleading news. *Journal of Computer-Mediated Communication* 30, 4 (06 2025), zmaf009. doi:10.1093/jcmc/zmaf009

[3] Andrew M. Guess, Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15536–15545. doi:10.1073/pnas.1920498117

[4] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. International Joint Conferences on Artificial Intelligence Organization, 8048–8057. doi:10.24963/ijcai.2024/890

[5] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206. doi:10.1162/tacl_a_00454

[6] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. doi:10.1145/3703155

[7] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. doi:10.1126/science.aao2998

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[9] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2356–2362. doi:10.1145/3404835.3463238

[10] Chang Lu, Bo Hu, Qiang Li, Chao Bi, and Xing-Da Ju. 2023. Psychological Inoculation for Credibility Assessment, Sharing Intention, and Discernment of Misinformation: Systematic Review and Meta-Analysis. *J Med Internet Res* 25 (29 Aug 2023), e49255. doi:10.2196/49255

[11] Miriam J. Metzger and Andrew J. Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics* 59 (2013), 210–220. doi:10.1016/j.pragma.2013.07.012

[12] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4551–4558. doi:10.24963/ijcai.2021/619

[13] Katherine Ognyanova, David Lazer, Ronald E. Robertson, and Christo Wilson. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review* (June 2020). doi:10.37016/mr-2020-024

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410

[15] Lisa Scharrer, Marc Stadtler, and Rainer Bromme. 2019. Judging scientific information: Does source evaluation prevent the seductive effect of text easiness? *Learning and Instruction* 63 (2019), 101215. doi:10.1016/j.learninstruc.2019.101215

[16] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. doi:10.18653/v1/N18-1074

[17] Pramukh Nanjundaswamy Vasist, Debashis Chatterjee, and Satish Krishnan. 2023. The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative. *Information Systems Frontiers* 26, 2 (April 2023), 663–688. doi:10.1007/s10796-023-10390-w

[18] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151. doi:10.1126/science.aap9559

[19] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 422–426. doi:10.18653/v1/P17-2067

[20] Sam Wineburg and Sarah McGrew. 2019. Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information. *Teachers College Record* 121, 11 (2019), 1–40. doi:10.1177/016146811912101102

[21] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversations. In *First Conference on Language Modeling*. https://openreview.net/forum?id=BAakY1hNKS

[22] Dake Zhang and Ronak Pradeep. 2023. ReadProbe: A Demo of Retrieval-Enhanced Large Language Models to Support Lateral Reading. arXiv:2306.07875 [cs.IR] https://arxiv.org/abs/2306.07875

[23] Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. 2024. Overview of the TREC 2024 Lateral Reading Track. In *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024) (NIST Special Publication)*. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec33/papers/Overview_lateral.pdf

[24] Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. 2025. Overview of the TREC 2025 DRAGUN Track: Detection, Retrieval, and Augmented Generation for Understanding News (Notebook Version). In *The Thirty-Fourth Text REtrieval Conference Proceedings (TREC 2025) (NIST Special Publication)*. National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec34/papers/Overview_dragun.pdf